TU Dortmund

Applied Bayesian Data Analysis

Project : Liver Cirrhosis Bayesian Data Analysis

Lecturers:

Prof. Dr. Katja Ickstadt

Prof. Dr. Paul Burkner

Group number: 15

Group members: Montasir Hasan Chowdhury, Md Emon Parvez, Mohaiminul Islam

March 17, 2024

Contents

1 Introduction

Liver cirrhosis remains a critical health issue worldwide, posing significant challenges in diagnosis, management, and treatment. With its multifactorial etiology and complex pathophysiology, understanding the factors influencing the advancement of liver cirrhosis is paramount for developing effective interventions. In this project, titled "Liver Cirrhosis Bayesian Data Analysis," we aim to explore the diverse array of factors impacting the progression of liver cirrhosis using advanced Bayesian statistical methods.

The primary research question guiding this investigation is: "How do different factors influence the advancement of liver cirrhosis?" Through rigorous analysis and interpretation of comprehensive datasets encompassing clinical, demographic, and laboratory parameters, we seek to unravel the intricate interplay between various factors and the progression of liver cirrhosis.

Our hypothesis posits that there exists a significant effect of several key factors on the status of liver cirrhosis patients. Specifically, we anticipate that factors such as drug usage, age, sex, bilirubin levels, triglycerides levels, platelet count, alkaline phosphatase levels, among others, will exert discernible influences on the progression and severity of liver cirrhosis. By investigating these factors within a Bayesian framework, we aim to not only quantify their individual effects but also elucidate potential interactions and synergistic relationships that may exist among them.

Bayesian data analysis offers a powerful approach for integrating prior knowledge, observational data, and expert judgment to generate robust insights and inform evidence-based decision-making. By leveraging the flexibility and interpretability of Bayesian methodologies, we aim to gain a deeper understanding of the dynamic processes underlying liver cirrhosis progression and identify novel avenues for personalized intervention and management.

Furthermore, this case study underscores the importance of interdisciplinary collaboration between clinicians, statisticians, and data scientists in tackling complex medical challenges. By synergizing clinical expertise with advanced statistical techniques, we aspire to contribute to the collective effort of advancing our understanding of liver cirrhosis and improving patient outcomes.

In the initial stages of the project, exploratory data analysis such as violin plots and correlation matrix will be employed to provide a visual understanding of the relationships between the variables. Subsequently, a Bayesian logistic regression model will be

constructed. In regard to bayesian logistic regression model, priors, model fitting, model fitting, model diagnostics, model evaluation on test data, model specifications, population level effects, group level effets, convergence diagnostics and model comparison will be done. Furthermore, sensitivity and specificity analysis will be done on the different generated model.

2 Problem statement

2.1 Description of the Data Set

The dataset utilized in this project has been meticulously collected from a reputable website called Kaggle, dataset title "Cirrhosis Patient Survival Prediction" (JoeBeach-Capital).

The dataset consists of 418 observations with 17 clinical attributes utilized for forecasting the survival outcomes of patients with liver conditions. These outcomes are categorized into 3 classes. Class 0: (C - censored), Class 1 (D - death), Class 2 = (CL - censored due to liver transplantation). Among the 17 clinical features, there are 5 integer variables which are Age, Cholesterol(serum cholesterol), Copper(urine copper), Triglycerides, Platelets(platelets per cube), 7 categorical variables which are Drug(type of drug), Sex(male or female), Ascites(presence of ascites), Hepatomegaly(presence of hepatomegaly), Spiders(presence of spiders), Edema(presence of edema), Stage(histologic stage of disease) and 5 continuous variables which are Bilirubin(serum bilirubin), Albumin, Alk_Phos(alkaline phosphatase), SGOT, Prothrombin(prothrombin time). There is 1 target variable that is Status which is a categorical variable and it represents the status of the patients.

2.2 Project Objective

The objective of this report is to investigate the influence of various covariates on the advancement of liver cirrhosis using Bayesian logistic regression models. Specifically, we aim to:

1.Implement three Bayesian logistic regression models, varying prior specifications and considering group variables.

2.Evaluate the performance of these models using the leave-one-out (LOO) cross-validation technique.

3.Assess the specificity and sensitivity of each model on test data to understand their predictive capabilities.

4.Provide insights into the effectiveness of different model specifications in capturing the relationship between covariates and the advancement of liver cirrhosis.

By accomplishing these objectives, this report aims to contribute to the understanding of liver cirrhosis progression and provide insights into the potential predictors that can aid in its early detection and management.

3 Code Description

3.1 Load Data

The dataset "cirrhosis.csv" containing information about liver cirrhosis patients is loaded into the R environment.

3.2 Synthetic Data Generation and Evaluation

Synthetic data is generated using the syn() function from the synthpop package and a function *create_unique_values_table*() is defined to generate a summary table comparing unique values, mean, skewness, kurtosis, and maximum category for each feature between the original and synthetic datasets.

3.3 Data Preprocessing and Feature Generation

Various preprocessing steps are performed, including one-hot encoding ,standardization, and generating new features based on domain knowledge.

3.4 Explanatory Analysis

The distributions of numerical features across different classes of the response variable are visualized to gain insights into the data. Based on these insights, necessary manipulations are performed on the response variable.

3.5 Correlation Analysis

A correlation plot is created using "corrplot" to investigate multicollinearity among features.

3.6 Train-Test Split

The dataset is split into training and testing sets using the createDataPartition() function to train (70%) and evaluate the models (30%).

3.7 Bayesian Logistic Regression Model

Bayesian logistic regression models are fitted using the brm package on the training data. These models allow for the incorporation of uncertainty in parameter estimates.

3.8 Model Comparison

Models are compared using the LOO function (loo()) to assess their predictive performance to different prior specifications and model complexicity.

3.9 Model Diagnostics

Model diagnostics, such as MCMC trace plots, are performed to evaluate the goodness-of-fit and ensure the models adequately capture the data patterns.

3.10 Model Evaluation on Test Split

Model performance is evaluated on the test split to assess their ability to generalize to new, unseen data. Sensitivity and specificity metrics are calculated to quantify model accuracy.

This comprehensive code description outlines the entire process of conducting Bayesian logistic regression analysis on liver cirrhosis data, emphasizing key steps for academic reporting purposes.

4 Data Preprocessing

This step involves preparing the data for analysis. Various preprocessing techniques are applied, such as handling missing values, scaling numerical features, and encoding categorical variables. One-hot encoding is a method used to convert categorical variables into a numerical format that First, it converts the 'Sex' variable to binary, where 'M' is encoded as 1 and other values as 0. Similarly, it encodes the 'Drug' variable as binary, where 'D-penicillamine' is encoded as 1 and other values as 0. The 'Edema' variable is converted to binary, with 'Y' encoded as 1 and other values as 0. The dataset is then reordered based on the 'Drug' variable. Additionally, 'Ascites', 'Spiders', and 'Hepatomegaly' variables are encoded as binary, with 'Y' being 1 and 'N'as 0. The original dataset comprises only 418 observations, which are subject to missing values. Such limitations in data availability may compromise the robustness and predictive efficacy of models derived from it. In response, synthetic data is generated to augment the dataset, providing a larger sample size for analysis andn mitigating the impact of missing values. Specifically, a synthetic dataset of 5000 samples is created from the original data, thereby expanding the data pool for analysis.

To evaluate the effectiveness of synthetic data in replicating the characteristics of the original dataset, various statistical properties are compared between the two datasets. These properties include measures such as unique values, mean, skewness, kurtosis, and the most frequent category for each feature. By assessing how well these properties align between the original and synthetic datasets, we can determine the extent to which the synthetic data accurately captures the underlying distribution and structure of the original dataset. Furthermore, additional features such as Thrombocytopenia, $Elevated\_alk\_phos$, $Normal\_Copper$, $Normal\_Albumin$, DiagnosisDays,

Age_Group, *Symptom_Score*, *Bilirubin_Albumin*, *Liver_Function_Index*, *Risk_Score*, *Diag_Month*, and *Diag_Year* are derived from existing features based on domain knowledge, enriching the dataset with relevant information. Our subsequent analyses and model development are predicated on this synthesized dataset, which encapsulates a more comprehensive representation of the original data.

5 Statistical Methods

In this section, several statistics methods are introduced which are later used in our analysis. The software R (R Core Team, 2022), corrplot (Wei and Simko, 2021), ggplot (Wickham, 2016), bayesplot (Gabry and Bürkner, 2024) are used for data analysis and visualization. To do data manipulation and transformation dplyr (Wickham, 2011) and caret (Fox and Weisberg, 2019) packages are utilized. To merge the multiple plots gridextra (Auguie, 2017) has been used. To do Bayesian regression models, brms (Bürkner, 2017) has been used. For generating synthetic data, synthpop (Nowok et al., 2016) has been used. To do leave-one-out cross-validation, loo (Vehtari et al., 2024) has been used.

5.1 Statistical Visualization Methods

This subsection will list all the statistical visualization methods that we used. All of them were plotted using libraries mentioned above in R.

5.1.1 Violin Plot

A violin plot is a type of visualization that combines the features of a box plot and a kernel density plot to depict the distribution of a continuous variable. Like a box plot, it includes a box that represents the interquartile range (IQR) of the data, with a line inside indicating the median. Additionally, it incorporates a kernel density estimation (KDE) plot on each side of the box, showing the probability density of the data at different values. This is achieved by smoothing the data with a kernel function. The width of the violin at any given point represents the density of data at that value. Mathematically, the KDE is calculated using the formula:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Where $\hat{f}(x)$ is the estimated density at point x, n is the number of data points, h is the bandwidth parameter, x_i are the data points, and K is the kernel function. Popular kernel functions include Gaussian and Epanechnikov kernels. The bandwidth parameter h controls the smoothing level of the KDE and is crucial for determining the trade-off between bias and variance in the estimation (Hintze and Nelson, 1998).

5.2 Correlation Matrix

A correlation matrix is a table that displays the correlation coefficients between several variables. Correlation coefficients quantify the strength and direction of the linear relationship between two variables, ranging from -1 to 1. A positive correlation coefficient indicates a positive linear relationship, meaning that as one variable increases, the other tends to increase as well. Conversely, a negative correlation coefficient suggests a negative linear relationship, where as one variable increases, the other tends to decrease. A correlation coefficient of 0 indicates no linear relationship between the variables. Mathematically, the correlation coefficient between two variables X and Y can be calculated using the Pearson correlation formula:

$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Where r is the correlation coefficient, X_i and Y_i are individual data points, $\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively. The correlation matrix extends this concept to multiple variables, providing a comprehensive overview of the relationships between all pairs of variables in a dataset. Each cell in the matrix represents the correlation coefficient between the variables corresponding to the row and column (Heumann et al., 2016).

5.3 Bayesian Logistic Regression

Bayesian logistic regression is a statistical method used for modeling the relationship between a binary outcome variable and one or more predictor variables, while incorporating Bayesian inference principles. It extends traditional logistic regression by considering uncertainty in model parameters and making probabilistic statements about them (Gelman et al., 2013).

Let's delve deeper into Bayesian logistic regression:

5.3.1 Model Specification

In Bayesian logistic regression, the model specification involves defining the relationship between predictor variables X and the log-odds of the binary outcome Y using the logistic function. Mathematically, this relationship is represented as:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_k X_k)}}$$

Here, β_0, β_1, ..., β_k are the coefficients representing the impact of each predictor variable on the log-odds of the outcome. The logistic function transforms the linear combination of predictors and coefficients into probabilities, ensuring that the predicted probabilities lie between 0 and 1, which is suitable for binary classification problems. The coefficients are interpreted as the change in the log-odds of the outcome for a one-unit change in the corresponding predictor variable, holding all other variables constant (Gelman et al., 2013).

5.3.2 Prior Specification

In Bayesian logistic regression, prior specification involves assigning prior distributions to the model parameters, particularly the coefficients (β). These prior distributions represent our beliefs or prior knowledge about the plausible values of the coefficients before observing any data. A common approach is to assume normal priors for each coefficient:

$$\beta_i \sim N(\mu_i, \sigma_i^2)$$

Here, μ_i and σ_i^2 are the mean and variance parameters of the prior distribution for β_i, respectively. These parameters reflect our beliefs about the central tendency and variability of the coefficients. By choosing appropriate values for μ_i and σ_i^2, we can incorporate prior knowledge or assumptions about the coefficients into the model. For instance, if we believe that certain predictors have a larger effect on the outcome than others, we can assign larger variances to their corresponding coefficients. Similarly, we

can specify informative priors based on previous studies or domain expertise to constrain the possible values of the coefficients (Gelman et al., 2013).

5.3.3 Posterior Inference

In Bayesian logistic regression, posterior inference involves updating our beliefs about the model parameters, particularly the coefficients (β), based on the observed data. According to Bayes' theorem, the posterior distribution of the coefficients given the data (Y and X) is proportional to the product of the likelihood function and the prior distribution:

$$P(\beta|Y,X) \propto P(Y|X,\beta) \cdot P(\beta)$$

Here, $\propto P(Y|X,\beta)$ represents the likelihood function, which is the probability of observing the data given the coefficients. In logistic regression, this likelihood function is the product of individual probabilities for each observation i:

$$P(Y|X,\beta) = \prod_{i=1}^{n} \left(\frac{1}{1+e^{-(\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\ldots+\beta_k X_{ik})}} \right)^{Y_i} \left(1 - \frac{1}{1+e^{-(\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\ldots+\beta_k X_{ik})}} \right)^{1-Y_i}$$

We then multiply the likelihood by the prior distribution $P(\beta)$ to obtain the joint posterior distribution of the coefficients. The posterior distribution reflects our updated beliefs about the coefficients after observing the data. However, obtaining the exact form of the posterior distribution is often analytically intractable, especially for complex models or large datasets. Therefore, we use Markov Chain Monte Carlo (MCMC) methods, such as Gibbs sampling or the Metropolis-Hastings algorithm, to obtain samples from the posterior distribution. These samples represent plausible values for the coefficients given the observed data and our prior beliefs. By summarizing these samples, we can perform inference by estimating posterior means, standard deviations, credible intervals, and other quantities of interest, providing insights into the uncertainty of parameter estimates and making predictions about new data. Bayesian posterior inference provides a principled framework for incorporating uncertainty into statistical analysis and decision-making, particularly in situations with limited data or when prior knowledge is available (Gelman et al., 2013).

5.3.4 Sampling from the Posterior

Sampling from the posterior distribution is a key step in Bayesian inference, particularly in Bayesian logistic regression. Once we have specified the likelihood function $\propto P(Y|X, \beta)$ and the prior distribution $P(\beta)$, obtaining samples from the joint posterior distribution $P(\beta|Y, X)$ allows us to make inferences about the model parameters. However, exact analytical solutions for the posterior distribution are often intractable, especially for complex models. Therefore, we resort to Markov Chain Monte Carlo (MCMC) methods, such as Gibbs sampling or the Metropolis-Hastings algorithm, to generate samples from the posterior distribution. These methods iteratively sample from the joint posterior distribution, converging to the true posterior distribution over time. In Gibbs sampling, we iteratively sample from the conditional posterior distributions of each parameter while holding the other parameters fixed. In the Metropolis-Hastings algorithm, we propose new parameter values based on a proposal distribution and accept or reject them based on a acceptance probability determined by the ratio of the posterior densities of the proposed and current parameter values. Through this iterative process, we obtain a set of samples from the posterior distribution, which can be used for posterior inference, such as estimating posterior means, standard deviations, credible intervals, and predictive distributions. Sampling from the posterior distribution enables us to incorporate uncertainty into our parameter estimates and make probabilistic statements about the model parameters, enhancing the robustness and interpretability of our statistical analyses (Gelman et al., 2013).

5.3.5 Inference

In Bayesian logistic regression, inference involves summarizing the posterior distribution of the model parameters obtained through sampling from the posterior. These summaries provide insights into the uncertainty of parameter estimates and can be used to make predictions about new data. Common summaries include posterior means, standard deviations, credible intervals, and posterior predictive distributions.

The posterior mean ($\hat{\beta}$) provides a point estimate of the coefficients, representing the average value of each coefficient across the posterior distribution. It is calculated as the average of the sampled values from the posterior distribution:

$$\hat{\beta} = \frac{1}{N} \sum_{i=1}^{N} \beta^{(i)}$$

where N is the number of samples and $\beta^{(i)}$ represents the i-th sampled value of the coefficients. The posterior standard deviation quantifies the uncertainty of the parameter estimates and is calculated as the standard deviation of the sampled values from the posterior distribution:

$$\text{Posterior SD}(\beta) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\beta^{(i)} - \hat{\beta})^2}$$

Credible intervals provide a range of plausible values for the coefficients and are typically constructed from the sampled values of the posterior distribution. For example, a 95% credible interval for a coefficient β_i consists of the middle 95% of the sampled values for β_i, excluding the extreme tails.

Posterior predictive distributions allow us to make predictions about new data given the observed data and the estimated model parameters. These distributions represent the uncertainty in the outcome variable given the predictor variables and can be used for tasks such as classification and regression. They are calculated by averaging the predictions from each sampled set of model parameters:

$$P(Y^*|X^*, Y, X) = \frac{1}{N} \sum_{i=1}^{N} P(Y^*|X^*, \beta^{(i)})$$

where Y^* represents the predicted outcome variable for new data X^*, N is the number of samples, and $P(Y^*|X^*, \beta^{(i)})$ is the predicted probability of Y^* given X^* and the i-th sampled set of model parameters (Gelman et al., 2013).

5.4 Leave-One-Out Cross-Validation (LOOCV)

Leave-One-Out Cross-Validation (LOOCV) is a technique used to evaluate the performance of a predictive model by systematically training the model on all but one observation in the dataset and then testing it on the omitted observation. Mathematically, for each observation i, the model is trained on the dataset excluding i, denoted as $D^{(-i)}$, and used to predict the response variable for observation i, denoted as $\hat{y}_i^{(-i)}$. The error

e_i for each observation i is calculated as the difference between the true response y_i and the predicted response $\hat{y}_i^{(-i)}$. Finally, the overall performance of the model, denoted as $\text{Error}_{\text{LOOCV}}$, can be computed using a performance metric such as Mean Squared Error (MSE), which aggregates the squared errors across all observations:

$$\text{Error}_{\text{LOOCV}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{(-i)})^2$$

This formula represents the mean of the squared errors, where n is the number of observations in the dataset (Magnusson et al., 2020).

5.5 Sensitivity and Specificity Analysis

5.5.1 Sensitivity Analysis

Sensitivity analysis is a method used in financial modeling, engineering, and other fields to assess how changes in the input variables of a model affect the output or outcomes. It involves systematically varying the input parameters within a defined range and observing the corresponding changes in the output. This analysis helps in understanding the robustness and reliability of a model by identifying which input variables have the most significant impact on the results, thus allowing decision-makers to focus their attention on critical factors. Sensitivity analysis can take various forms, such as one-way sensitivity analysis, which examines the impact of changing one input variable while keeping others constant, or multi-way sensitivity analysis, which considers multiple input variables simultaneously. Additionally, it aids in risk management by highlighting areas of uncertainty and providing insights into potential scenarios under different conditions, ultimately supporting informed decision-making processes (Vidakovic, 2011).

5.5.2 Specificity Analysis

Specificity analysis, often used in the context of medical diagnostics, refers to a statistical measure that evaluates the ability of a test to correctly identify individuals without a particular condition or characteristic, also known as true negatives. It quantifies the proportion of negative cases that are correctly identified by the test out of all true negative cases. Specificity analysis is crucial for assessing the reliability and accuracy of diagnostic tests, especially in distinguishing between healthy individuals and those who

do not have a specific condition or disease. A high specificity indicates that the test is effective in ruling out the condition in healthy individuals, minimizing the occurrence of false positive results, which are cases where the test incorrectly indicates the presence of the condition in individuals who do not have it. Specificity analysis plays a vital role in evaluating the performance of diagnostic tests and aids clinicians in making accurate diagnostic decisions (Vidakovic, 2011).

6 Statistical Analysis

6.1 Exploratory Data Analysis

6.1.1 Distributions of Features

We have distributed each features into 3 classes. These 3 classes are Class 0: (C - censored), Class 1 (D - death), Class 2 = (CL - censored due to liver transplantation).

Feature: *N_Days*

The variable *N_Days* denotes the duration in days from registration to either death, transplantation, or the time of study analysis, whichever occurred earlier.

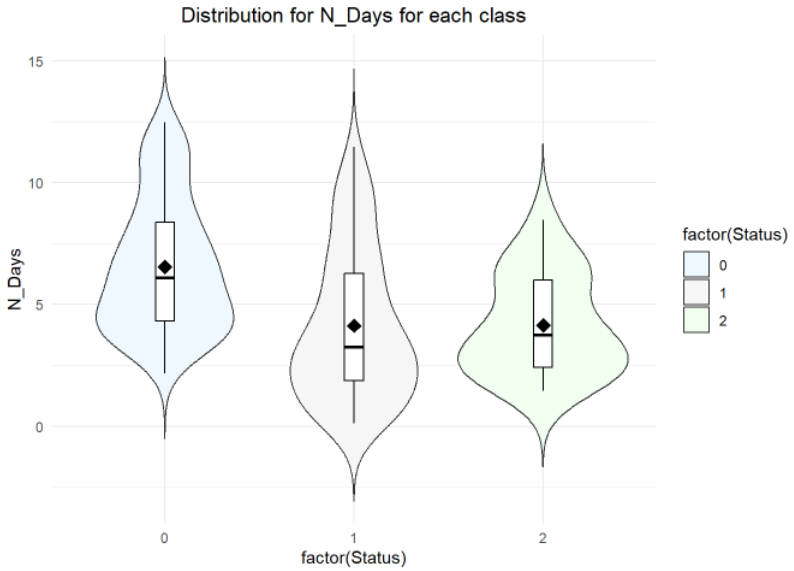

Figure 1: Distribution for *N_Days* for each class

Looking at Figure 1, it's evident that Class 1 shows a slightly lower distribution, mainly because a majority of them passed away before the study concluded. The pattern is similar for liver transplant recipients. On the other hand, Class 0 patients, who were censored, continued throughout the study, presumably because they either survived or were not lost to follow-up.

Feature: Age

Figure 2 reveals that the events predominantly happen in individuals aged 40 and above. Class 1 events, primarily deaths, are more common in older age groups. Conversely, most individuals in Class 2, who undergo liver transplants, do so before reaching the age of 50.

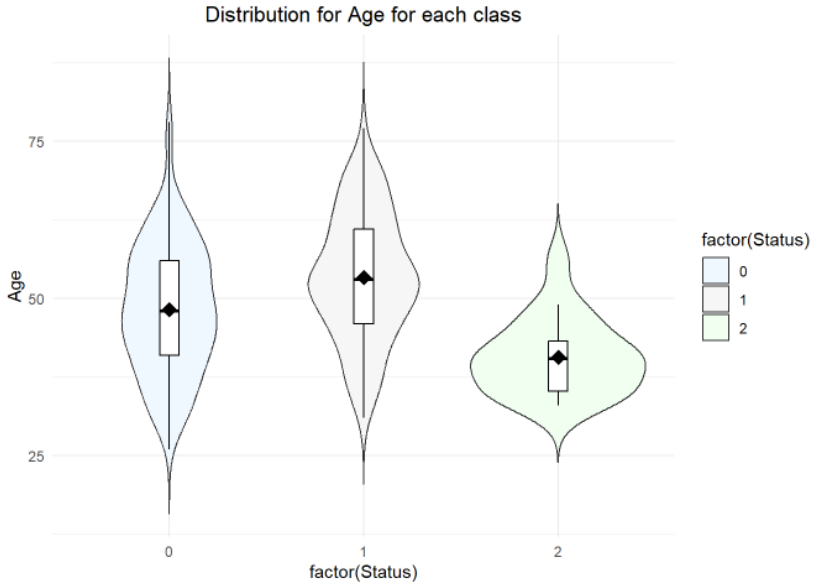

Figure 2: Distribution for Age for each class

Feature: Bilirubin

Higher levels of serum bilirubin in the blood correlate with increased liver complications.

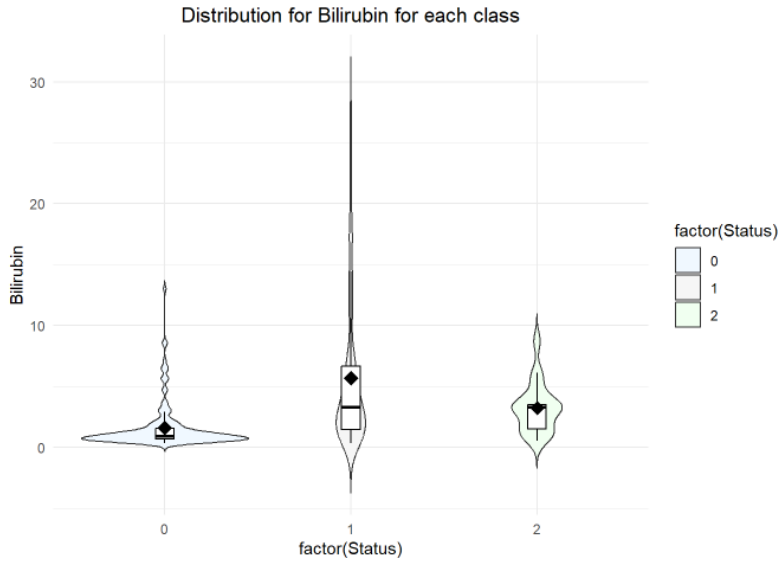

Figure 3: Distribution for Bilirubin for each class

Figure 3 suggests that individuals in Class 1 likely experienced severe liver damage, as indicated by their elevated bilirubin levels. In contrast, Class 0 individuals, who were censored, had lower bilirubin levels, indicating their condition was generally under control with less damage. Class 2 individuals also exhibited higher bilirubin levels.

Feature: Albumin

Decreased liver function leads to a reduction in albumin production.

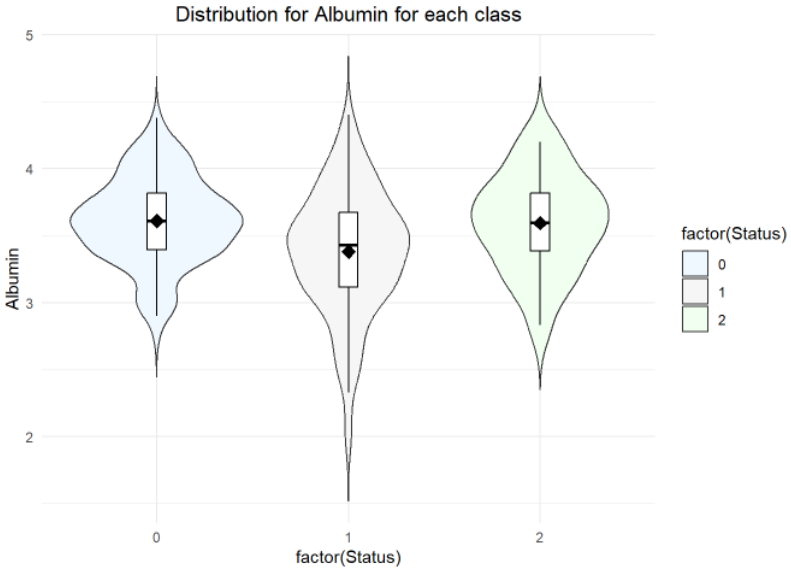

Figure 4: Distribution for Albumin for each class

Figure 4 indicates that Class 1 patients, who did not survive, experienced a decline in albumin levels compared to both Class 0 patients and those in Class 2.

Feature: Copper

Cirrhosis can disrupt copper metabolism, resulting in its buildup within the liver.

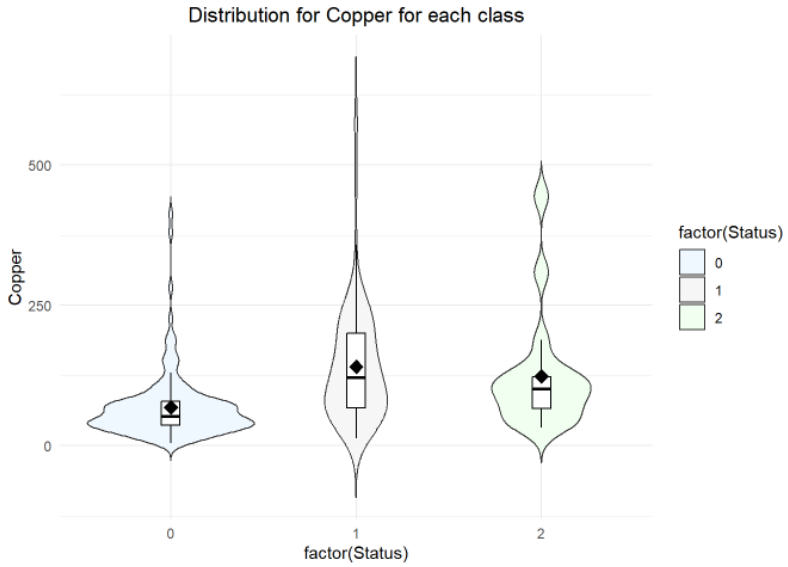

Figure 5: Distribution for Copper for each class

By examining the median copper levels in Class 1 patients who did not survive, as depicted in Figure 5.

Feature: *Alk_Phos* (Alkaline Phosphatase)

An increase in alkaline phosphatase levels can occur as a result of bile duct obstruction or cholestasis, which are frequent occurrences in cirrhosis.

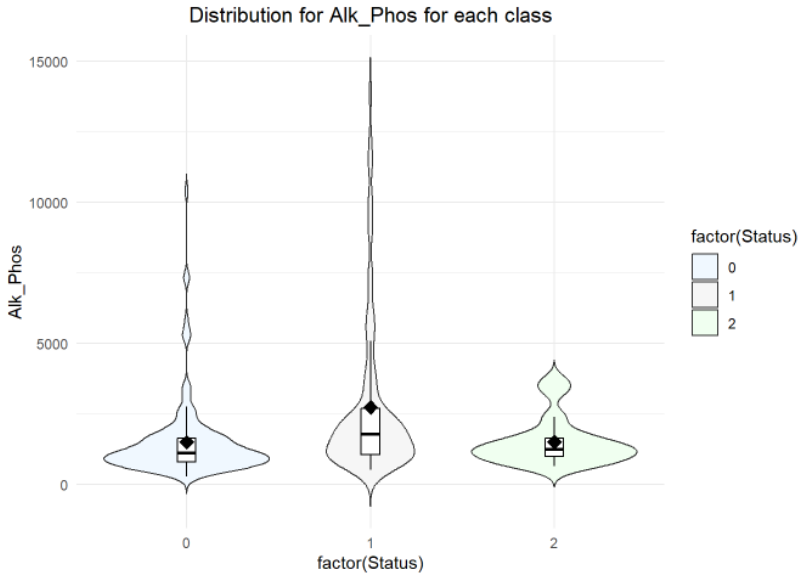

Figure 6: Distribution for *Alk_Phos* for each class

In Figure 6, we can see that the median alkaline phosphatase levels are notably elevated in both Class 1 and Class 2 patients.

Feature: Serum Glutamic Oxaloacetic Transaminase (SGOT)

Increased levels of Serum Glutamic Oxaloacetic Transaminase (SGOT) may signify damage to liver cells.

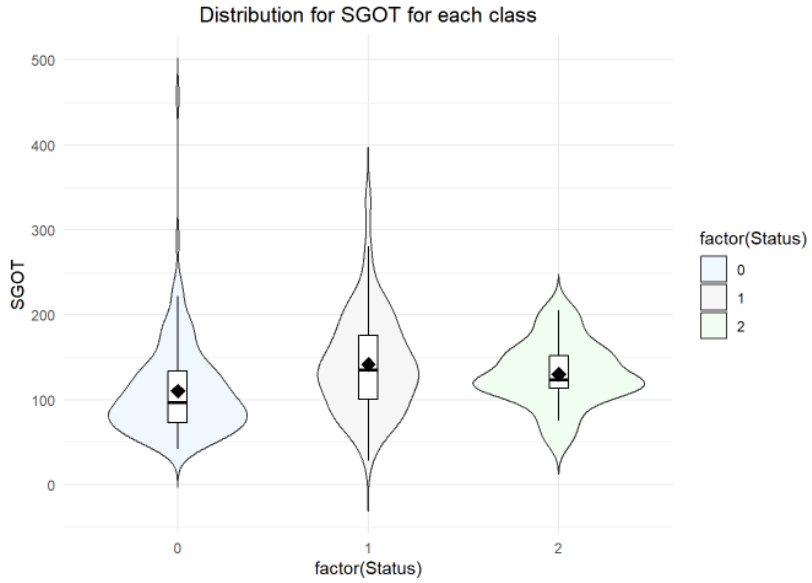

Figure 7: Distribution for SGOT for each class

In Figure 7, it's evident that the median SGOT levels are significantly elevated among patients in Class 1 and Class 2.

Feature: Triglycerides

Elevated triglyceride levels can result from impaired liver function.

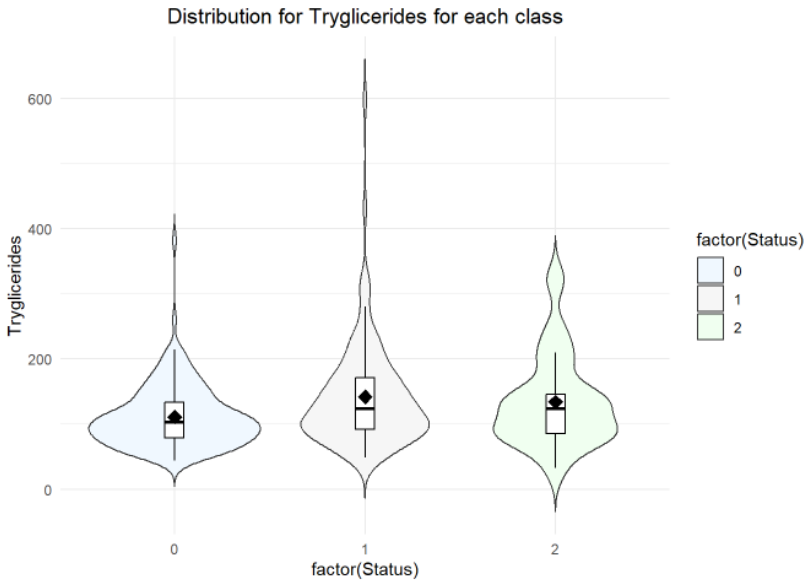

Figure 8: Distribution for Triglycerides for each class

In Figure 8, it's noticeable that the median triglyceride levels are slightly elevated among patients in Class 1 and Class 2.

Feature: Platelets

As cirrhosis advances, it is common for the platelet count to decrease.

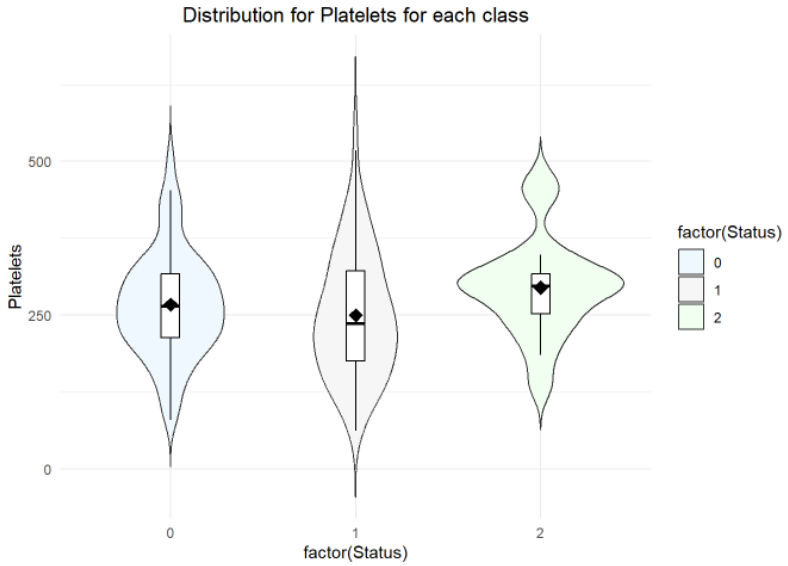

Figure 9: Distribution for Platelets for each class

Figure 9 illustrates that individuals in Class 1 who did not survive exhibit a significantly lower platelet count.

Feature: Prothrombin Time (PT)

Prothrombin time, a gauge of blood clotting, may be extended in cirrhosis owing to diminished production of clotting factors by the liver.

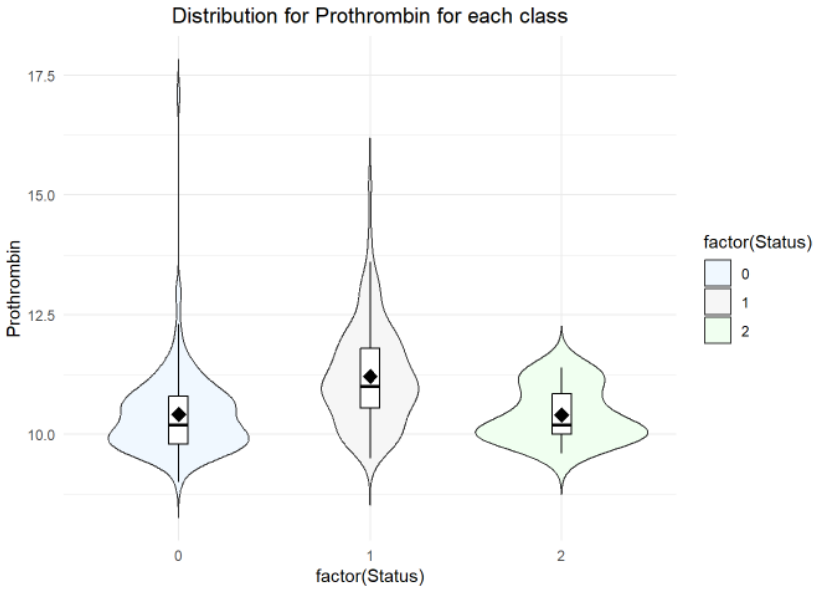

Figure 10: Distribution for Prothrombin for each class

Figure 10 indicates a notably elevated prothrombin time among individuals in Class 1 who did not survive.

Judging from the distributions of the mentioned features, it appears that patients in Class 1 exhibit distinct characteristics compared to those in Class 0 and Class 2. Hence, for the convenience of our analysis, we have consolidated Class C-censored (0) and Class Cl-censored due to liver transplant (2) into a single class labeled as "Survived" (0). Consequently, these merged classes now represent individuals .As a result, the response variable "status" has been transformed into a binary outcome with two distinct classes: "Survived" (0) and "Death" (1). This binary categorization allows us to effectively apply logistic regression modeling techniques to investigate the factors influencing the likelihood of mortality versus survival in patients with liver cirrhosis.

6.1.2 Correlation Matrix

Based on the correlation plot analysis in figure 11, several explanatory variables were found to exhibit high correlation, particularly among the engineered features. Specifically, the variables $Diag\_year$ and N_Days, $Liver\_function\_index$ and Alk_phos, and $Bilirubin\_Albumin$ and Copper demonstrated strong positive correlations among themselves. Conversely, $Risk\_Score$ and $Liver\_function\_index$, as well as $Liver\_function\_index$ and $Risk\_Score$, displayed notable negative correlations.

To address the issue of multicollinearity and avoid potential autocorrelation in the model, the following variables were excluded from the analysis: $Diagnosis\_Days$, $Bilirubin\_Albumin$, $Diag\_Year$, $Risk\_Score$, and $Liver\_Function\_Index$. This decision was made to enhance the interpretability and stability of the regression model, ensuring that the estimates of coefficients remain unbiased and reliable.

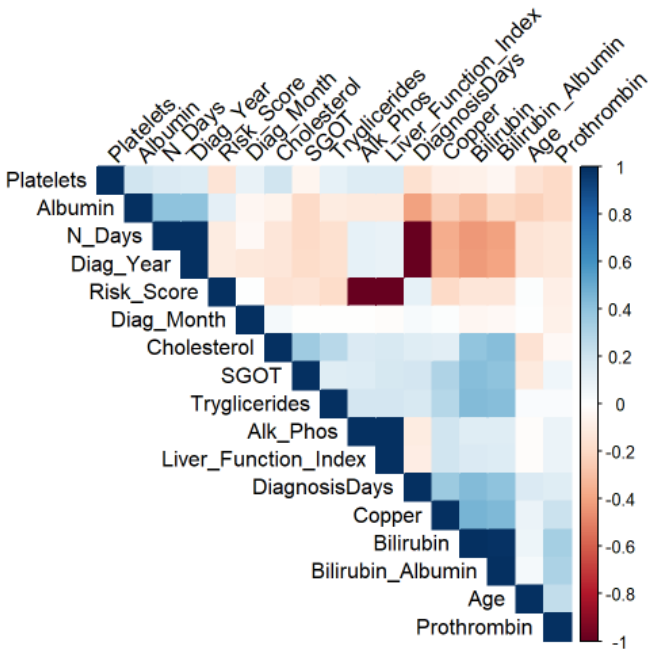

Figure 11: Correlation Matrix

6.2 Bayesian Logistic Regression Analysis

6.2.1 Priors

In our bayesian logistic regression analysis, we aimed to understand the factors influencing the odds of death compared to survival. However, we began our modeling process without any prior knowledge about the parameters. After exploring the dataset, we developed a general expectation that certain covariates such as age, cholesterol levels, and others may contribute to an increase in the odds of death. Conversely, we hypothesized that the use of D-penicillamine (Drug1) might lead to a decrease in the odds of death.

To incorporate these beliefs into our model, we employed different prior specifications. For most covariates, excluding Drug1, we utilized a skewed prior centered around 1 for the parameter θ. This reflects our strong belief that the odds of death are likely to be higher than those of survival. We quantified this expectation by setting θ to 1, where the resulting odds ($\exp(\theta) \approx 2.71$) indicates a heightened chance of death.

Conversely, for the Drug1 variable, we adopted a more specific prior. We selected -0.5 as the most plausible value for the parameter, indicative of our belief that the odds of

death may decrease with the use of penicillin. By setting θ to -0.5, we implied that the odds ($\exp(\theta) \approx 0.61$) would be lower, suggesting a reduced likelihood of death.

To contrast these assumptions, we considered an alternative prior distribution centered around 0 for all covariates, reflecting uncertainty about the direction of their effects on the odds of death. Here, $\exp(\theta) = 1$ implies no discernible increase or decrease in the odds of death relative to survival. This approach allowed us to explore a range of possibilities and better understand the impact of our prior beliefs on the logistic regression model.

6.2.2 Model Specifications

For model 1, encompasses a logistic regression model fitted to the training dataset, aimed at predicting the binary response variable "status," where death is encoded as 1 and survival as 0. This regression is conducted on a set of explanatory variables. Moreover, to accommodate potential group effects, random intercepts are incorporated for the variables "*Age_Group*", "Stage" and "*Symptom_Score*". To encapsulate prior beliefs about the relationship between the covariates and the odds of death, two distinct types of prior distributions are employed. Firstly, a student's t-distribution with parameters df=4, mean=1, and scale=2 is assigned to all covariates except "Drug". Secondly, a normal distribution with a mean of -0.5 and a variance of 0.25 is assigned specifically to the covariate "Drug". In pursuit of posterior inference, the model undergoes a Markov Chain Monte Carlo (MCMC) sampling process. Specifically, the total number of iterations is set at 4000, with a warmup period comprising 2000 iterations. This approach aims to achieve convergence and facilitate accurate estimation of model parameters.

For model 2, comprises a logistic regression model applied to the same training dataset, with the response variable "Status" regressed on all available explanatory variables. Random intercepts are included for the grouping variables "*Age_Group*", "Stage" and "*Symptom_Score*" indicating a consideration for potential group effects. In contrast to Model 1, We considered an alternative prior distribution centered around 0 for all covariates, reflecting uncertainty about the direction of their effects on the odds of death. Specifically, a student's t-distribution with parameters df=3, mean=0, and scale=2 is employed to encapsulate prior beliefs about the association between the covariates and the likelihood of death. For this model, Control parameters are adjusted to ensure convergence, with the *adapt_delta* parameter set to 0.98. The total number of iterations is 4000, with a warmup period comprising 2000 iterations.

For model 3, model 3 closely resembles model 1, featuring a logistic regression model applied to the same training data to predict the binary response variable "Status". Both models utilize all available explanatory variables except the fact that the consideration of group effects, such as random intercepts for the grouping variables "*Age_Group*" and "Stage" is ignored. Despite the difference in model complexity, both model 1 and model 3 adopt similar prior specifications for parameter estimation. The total number of iterations is set at 4000, with a warmup period comprising 2000 iterations, ensuring convergence and accurate estimation of model parameter

6.2.3 Model Fitting

Bayesian logistic regression models are fitted to the preprocessed data using the 'brms' package. Bayesian regression allows for uncertainty in model parameters to be quantified, providing more robust estimates compared to traditional frequentist approaches. Different prior specifications for model parameters are considered, reflecting varying levels of prior belief or knowledge about the relationships between predictors and the outcome variable.

Interpretation of Model Coefficients:

Population Level Effects:

Table 1: Population Level Effects

Coefficient	Model 1 Estimate	Model 2 Estimate	Model 3 Estimate
Intercept	-4.93	-2.50	-4.96
N_Days	-0.81	-0.75	-0.80
Drug1	-0.22	0.98	-0.22
Age	0.21	0.18	0.22
Sex1	0.77	0.58	0.74
Ascites1	1.04	0.60	0.90
Hepatomegaly1	-1.34	-1.39	-1.27
Spiders1	0.72	0.57	0.71
Edema1	6.31	8.38	5.94
Bilirubin	1.45	1.59	1.39
Cholesterol	0.06	-0.04	0.05
Albumin	0.10	0.05	0.10
Copper	-0.12	-0.14	-0.13
Alk_Phos	1.05	1.04	1.03
SGOT	0.60	0.65	0.59
Tryglicerides	0.48	0.51	0.47
Platelets	0.28	0.22	0.27
Prothrombin	0.90	0.94	0.85
Stage2	2.02	1.19	2.34
Stage3	1.52	0.80	1.64
Stage4	1.57	0.78	1.73
thrombocytopenia1	1.17	0.66	1.16
elevated_alk_phos	0.99	-0.01	1.00
*normal_copper*1	-0.51	-0.69	-0.48
*normal_albumin*1	0.22	0.20	0.19
*Age_Group*1	0.17	-0.66	-0.10
*Age_Group*2	2.09	1.30	2.42
*Age_Group*3	0.94	-0.03	0.94
*Symptom_Score*1	1.56	1.06	1.65
*Symptom_Score*2	1.67	1.25	1.76
*Symptom_Score*3	-1.10	-1.79	-1.42
Diag_Month	0.86	0.90	0.84

Interpretation:

Intercept:

In Model 1, the estimated baseline log-odds of death when all other predictor variables are zero is approximately –4.93. For Model 2 and Model 3, these estimates are -2.50 and -4.96, respectively, with a high degree of uncertainty. This typically suggests that the probability of death decreases when other covariates' effects are ignored.

N_Days (in years):

The longer the reporting days, the lower the chance of death, indicating that earlier detection of Liver Cirrhosis decreases the odds of death. Each model suggests that for each additional N-day, the odds of death decrease by approximately exp(0.81), exp(0.75), and exp(0.80), respectively, for Model 1, Model 2, and Model 3.

Drug1:

In Model 1 and Model 3, the use of D-penicillamine decreases the chance of death compared to a placebo, with odds of exp(0.22). However, Model 2 predicts a higher chance of death for the use of D-penicillium, i.e., exp(0.98).

Age:

Older age is associated with a higher chance of death due to liver cirrhosis. Model 1, Model 2, and Model 3 estimate increases in the odds of death by approximately exp(0.21), exp(0.18), and exp(0.22), respectively.

Sex1:

The probability of death is higher in males compared to females due to liver cirrhosis. Model 1, Model 2, and Model 3 show odds of exp(0.77), exp(0.58), and exp(0.74), respectively.

Ascites1:

Presence of ascites increases the odds of death, with exp(1.04) in Model 1 and exp(0.90) in both Model 2 and Model 3. Hepatomegaly1: Presence of hepatomegaly significantly decreases the odds of death, with exp(1.34) in Model 1, exp(1.39) in Model 2, and exp(1.27) in Model 3.

Spiders1:

Presence of spiders increases the chance of death, with the log odds of death increasing by approximately the same amount in each model, approximately exp(0.72).

Edema1:

Presence of edema significantly increases the odds of death, with exp(6.31) in Model 1, exp(8.38) in Model 2, and exp(5.94) in Model 3.

Bilirubin:

Increased bilirubin levels have adverse effects on liver cirrhosis, with each unit increase resulting in a increase in the odds of death, with exp(1.45), exp(1.59), and exp(1.39) for Model 1, Model 2, and Model 3, respectively.

Cholesterol:

Each unit increase in cholesterol is associated with an increase in the odds of death in Model 1 and Model 3, but Model 2 indicates the opposite relationship, with exp(-0.04).

Albumin:

Each unit increase in albumin results in an increase in the odds of death for all models, with exp(0.10).

Copper:

Each unit increase in copper leads to a decrease in the odds of death, with approximate decreases of exp(0.12), exp(0.14), and exp(0.13) for Model 1, Model 2, and Model 3, respectively.

Alkaline phosphatase:

Each unit increase in alkaline phosphatase is associated with an increase in the odds of death, with exp(1.05) for Model 1, exp(1.04) for Model 2, and exp(1.03) for Model 3.

SGOT:

Serum Glutamic-Oxaloacetic Transaminase is a protein made by liver cells. Each unit increase is associated with a odds of death increase by approximately exp(0.60) for Model 1, exp(0.65) for Model 2, and exp(0.59) for Model 3. Significant

Triglycerides:

Each unit increase in triglycerides results in an increase in the odds of death approximate increases of exp(0.48), exp(0.51), and exp(0.47) for Model 1, Model 2, and Model 3 respectively.

Platelets:

Each unit increase in platelets leads to an increase in the odds of death, with approximate increases of exp(0.28), exp(0.22), and exp(0.27) for Model 1, Model 2, and Model 3, respectively.

Prothrombin:

Each unit increase in prothrombin is associated with an increase in the odds of death, with exp(0.90) for Model 1, exp(0.94) for Model 2, and exp(0.85) for Model 3.

Thrombocytopenia:

Presence of thrombocytopenia increases the odds of death, with exp(1.17) for Model 1, exp(0.66) for Model 2, and exp(1.16) for Model 3.

Elevated alkaline phosphatase:

Presence of elevated alkaline phosphatase significantly increases the odds of death, with exp(0.99) for Model 1, exp(1.00) for Model 2, and exp(1.00) for Model 3.

Normal copper levels:

Having normal copper levels decreases the odds of death compared to higher copper levels, with approximate decreases of exp(0.51), exp(0.69), and exp(0.48) for Model 1, Model 2, and Model 3, respectively.

Diag_month:

As the diagnosis month increases, so does the probability of death, indicating that the probability of death increases when the patient is diagnosed with liver cirrhosis for a longer period of time. For Model 1, the odds of death increase with exp(0.86), and for Model 2 and Model 3, they increase by exp(0.84) with each month increased, with statistical significance.

Group Level Effects:

The analysis of group-level effects indicates significant variations in the baseline chance of death, or odds of death, within each group for each model.

Model 1:

The estimated standard deviations are 1.42, 1.36, and 1.43 for *Age_Group*, Stage, and *Symptom_score*, respectively. The 95% credible intervals range from approximately 0.05 to 4.73, indicating a high level of uncertainty in the precise level of variability but confirming the existence of variability across each group. For example, *Age_Group* 2 shows a higher probability of death, with odds exp(2.09), compared to the reference age

group (*Age_Group* 0). In contrast, *Age_Group* 1 and *Age_Group* 3 exhibit different and comparatively lesser effects.

Model 2:

The estimated standard deviations are 1.40, 1.66, and 1.59 for *Age_Group*, Stage, and *Symptom_score*, respectively. The 95% credible intervals range from approximately 0.05 to 5.05, indicating higher uncertainty compared to Model 1 in the precise level of variability but still confirming the presence of variability across each group.

Influence of Prior and Model Complexity:

The complexity of the model and the choice of prior can indeed influence the estimation of coefficients. This phenomenon is evident in our analysis, particularly with regards to the Drug variable and the Cholesterol covariate. In Model 1 and Model 3, where we incorporate our prior belief of decreasing the odds of death due to D-penicillamine (with and without group effect), the coefficient of the Drug variable has a negative log odds estimation (-0.22). However, in Model 2, where we didn't specify any prior direction with the Drug variable, we observe a positive estimation for the coefficient (0.96), indicating that the odds of death increase with the use of D-penicillamine. A similar phenomenon is observed with the Cholesterol covariate, where Model 2 shows an opposite direction in the development of liver cirrhosis compared to the other two models. Thus, it becomes evident that model complexity and the proper choice of prior are necessary for optimal model estimation and interpretation of results.

6.2.4 Model Diagnostics

Diagnostic checks, such as trace plots is conducted to assess the validity and reliability of the fitted models. Trace plots visualize the sampling process of the Markov chain Monte Carlo (MCMC) algorithm used in Bayesian inference, helping to diagnose issues like poor mixing or lack of convergence.

Convergence Diagnostics:

Model 1:

Gelman-Rubin statistic (Rhat): The Rhat values for all parameters are 1, indicating good convergence among chains.

Effective Sample Size (ESS): Both bulk ESS and tail ESS are sufficiently high for all parameters, suggesting effective sampling.

Bulk ESS: The bulk ESS values range from several thousand to over ten thousand, indicating adequate sampling from the bulk of the posterior distribution.

Tail ESS: The tail ESS values are also high, indicating that even the tails of the posterior distribution are well-sampled.

Adapt Delta: The adapt delta value used for the NUTS algorithm was 0.98.

Posterior Draws: The model was run with 4 chains, each with iter = 4000, resulting in a total of 8000 post-warmup draws.

Number of Chains: 4

Total Draws: 8000

Nuts algorithm: The NUTS algorithm was used for sampling.

Model 2:

Gelman-Rubin statistic (Rhat): The Rhat values for all parameters are 1, indicating good convergence among chains.

Effective Sample Size (ESS): Both bulk ESS and tail ESS are sufficiently high for all parameters, suggesting effective sampling.

Bulk ESS: The bulk ESS values range from several thousand to over ten thousand, indicating adequate sampling from the bulk of the posterior distribution.

Tail ESS: The tail ESS values are also high, indicating that even the tails of the posterior distribution are well-sampled.

Adapt Delta: The adapt delta value used for the NUTS algorithm was 0.98.

Posterior Draws: The model was run with 4 chains, each with iter = 4000, resulting in a total of 8000 post-warmup draws.

Number of Chains: 4

Total Draws: 8000

Nuts algorithm: The NUTS algorithm was used for sampling.

Model 3:

Gelman-Rubin statistic (Rhat): The Rhat values for all parameters are 1, indicating good convergence among chains.

Effective Sample Size (ESS): Both bulk ESS and tail ESS are sufficiently high for all parameters, suggesting effective sampling.

Bulk ESS: The bulk ESS values range from several hundred to several thousand, indicating adequate sampling from the bulk of the posterior distribution.

Tail ESS: The tail ESS values are also high, indicating that even the tails of the posterior distribution are well-sampled.

Adapt Delta: No adapt delta value is provided for Model 3.

Posterior Draws: The model was run with 4 chains, each with iter = 2000, resulting in a total of 4000 post-warmup draws.

Number of Chains: 4

Total Draws: 4000

Nuts algorithm: The NUTS algorithm was used for sampling.

All three models display good convergence: Rhat values are close to 1, indicating convergence among chains. Effective sample sizes (ESS) are high, reflecting efficient sampling. The adapt delta value for Models 1 and 2 is 0.98. Model 3 lacks an adapt delta value but still converges well. Total draws are 8000 for Models 1 and 2 and 4000 for Model 3.

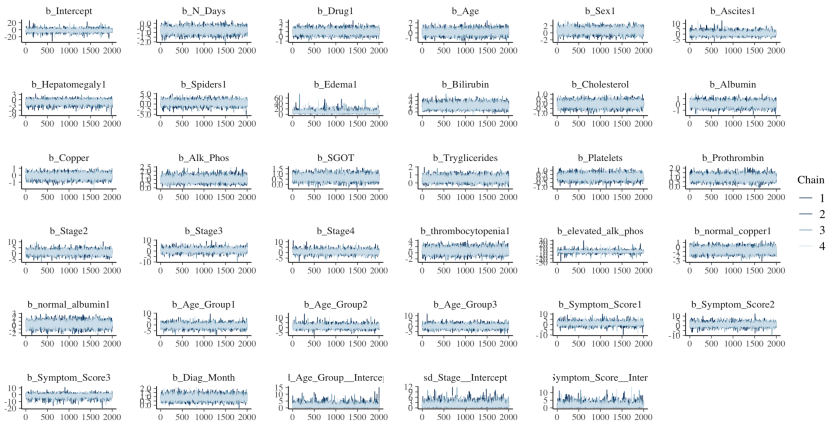

Figure 12: Model 1 Trace Plot

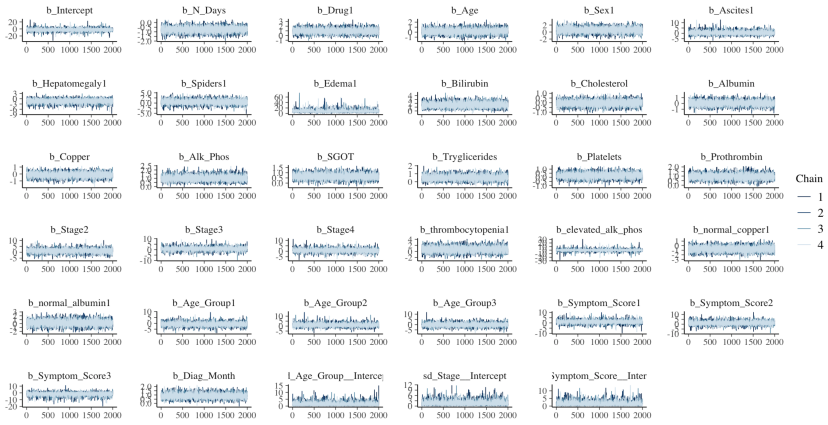

Figure 13: Model 2 Trace Plot

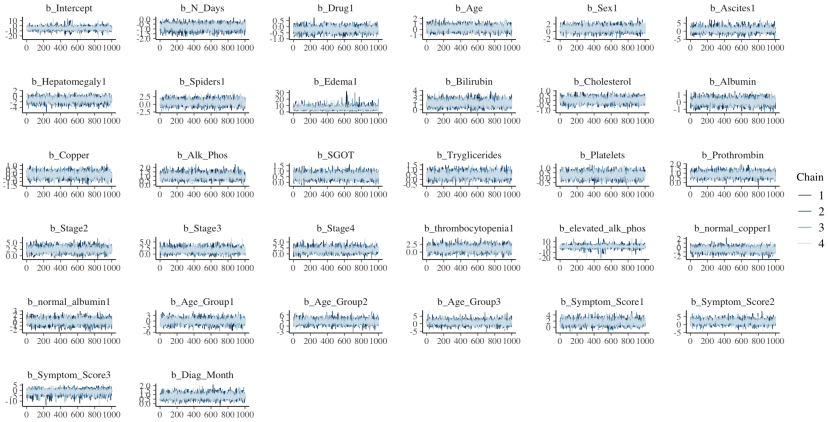

Figure 14: Model 3 Trace Plot

While examining the trace plots in figure 12, figure 13 and figure 14, it can be seen that none of the four chains overlap, indicating independent and thorough exploration of the parameter space by each chain. By paying attention to Gelman-Rubin statistics which is 1, we can confirm a good convergence. Overall, this trace plot demonstrates reliable MCMC sampling and robust posterior inference.

6.2.5 Model Comparison

Models are compared using leave-one-out cross-validation (LOO), a rigorous validation technique that assesses the predictive accuracy and generalization ability of the models. LOO involves fitting the model multiple times, each time leaving out one observation from the training set and evaluating the model's performance on the left-out observation.

This process provides a more reliable estimate of how well the model will perform on new, unseen data.

The Pareto k diagnostic values provide insights into the stability of estimates in each model.

Model1:

Good (79.4%): 154 observations. This category indicates that the majority of observations have stable estimates, suggesting robust model performance.

Ok (11.9%): 23 observations. These observations have acceptable stability but may warrant some caution.

Bad (8.8%): 17 observations. A small proportion of observations fall into this category, indicating fewer stable estimates that might require further investigation.

Very Bad (0.0%): 0 observations. There are no observations with highly unstable estimates in this model.

Model2:

Good (84.5%): 164 observations. The highest percentage of observations fall into this category, indicating strong stability and robust model performance.

Ok (10.3%): 20 observations. A small percentage of observations have acceptable stability but may require some attention.

Bad (4.6%): 9 observations. A few observations exhibit fewer stable estimates, suggesting potential issues in these cases.

Very Bad (0.5%): 1 observation. There is only one observation with highly unstable estimates, which might be an outlier.

Model3:

Good (84.0%): 163 observations. Similar to Model2, most observations have stable estimates, indicating robust model performance.

Ok (10.8%): 21 observations. These observations have acceptable stability but may require some scrutiny.

Bad (3.6%): 7 observations. A small proportion of observations exhibit fewer stable estimates, suggesting potential issues in these cases.

Very Bad (1.5%): 3 observations. There are a few observations with highly unstable estimates, indicating potential outliers or problematic cases.

Model Comparisons:

Model1 vs. Model2:

$elpd\_diff$: Model1's expected log pointwise predictive density ($elpd\_loo$) is 4.2 units lower than Model2. This suggests that Model2 has better predictive performance than Model1.

se_diff: The standard error of the difference in $elpd\_loo$ between Model1 and Model2 is 3.0, indicating the uncertainty in this difference.

Model2 vs. Model3:

$elpd\_diff$: Model2's $elpd\_loo$ is 3.2 units higher than Model3. This indicates that Model2 performs slightly better in terms of predictive performance compared to Model3.

se_diff: The standard error of the difference in $elpd\_loo$ between Model2 and Model3 is 3.0, implying some uncertainty in this difference.

The comparison of three Bayesian models, Model1, Model2, and Model3, reveals distinctive performance characteristics. Model1 exhibits the weakest predictive performance, with the highest expected log pointwise predictive density ($elpd\_loo$) of -97.7 and a corresponding LOO Information Criterion (looic) of 195.4, indicating poorer model fit compared to the others. Conversely, Model2 emerges as the top performer, boasting the lowest $elpd\_loo$ of -93.5 and the lowest looic of 186.9, signifying superior predictive accuracy and model fit. Moreover, Model2 demonstrates stable estimates with the highest percentage of observations falling into the "good" category according to the Pareto k diagnostic values. Model3 falls between Model1 and Model2 in terms of performance metrics, with $elpd\_loo$ of -96.7 and looic of 193.3, indicating better predictive performance than Model1 but slightly trailing behind Model2. Ultimately, Model2 stands out as the optimal choice among the three models, offering the best balance of predictive accuracy, model complexity, and stability of estimates.

6.2.6 Model Evaluation on Test Data

Finally, the performance of the fitted models is evaluated on a separate test dataset. Metrics such as sensitivity and specificity are calculated to assess the model's ability to correctly identify cases of cirrhosis. Sensitivity measures the proportion of true positives correctly identified by the model, while specificity measures the proportion of true negatives correctly identified. These metrics provide valuable insights into the practical

utility of the models in real-world healthcare settings. By following these steps and utilizing the specified libraries, the code aims to provide a comprehensive analysis of the cirrhosis dataset, yielding valuable insights for healthcare analytics and decision-making.

Sensitivity and Specificity Analysis:

Model1:

Sensitivity (True Positive Rate): 0.6875. Model1 correctly identifies approximately 68.75% of the actual positive cases. This indicates its ability to effectively detect instances of the positive class.

Specificity (True Negative Rate): 0.78. Model1 demonstrates a specificity of 0.78, meaning it correctly identifies around 78% of the actual negative cases. This suggests its capacity to accurately distinguish instances of the negative class.

Model2:

Sensitivity (True Positive Rate): 0.6471. Model2 achieves a sensitivity of approximately 64.71%, indicating its capability to correctly identify about 64.71% of the actual positive cases.

Specificity (True Negative Rate): 0.7708. With a specificity of 0.7708, Model2 accurately identifies around 77.08% of the actual negative cases. This highlights its effectiveness in discerning instances of the negative class.

Model3:

Sensitivity (True Positive Rate): 0.6563. Model3 demonstrates a sensitivity of approximately 65.63%, suggesting its ability to correctly identify about 65.63% of the actual positive cases.

Specificity (True Negative Rate): 0.76. With a specificity of 0.76, Model3 accurately identifies around 76% of the actual negative cases. This indicates its proficiency in distinguishing instances of the negative class.

The analysis of sensitivity and specificity across three models, Model1, Model2, and Model3, reveals their respective capabilities in binary classification tasks. While all models demonstrate relatively high specificity, indicating their ability to accurately identify instances of the negative class, their sensitivity values vary slightly. Model1 exhibits the highest sensitivity at 0.6875, followed closely by Model3 (0.6563) and Model2 (0.6471). This indicates Model1's superior performance in correctly identifying instances of the

positive class. However, all models maintain a balance between sensitivity and specificity, suggesting their effectiveness in classification tasks.

6.3 Predictive Performance Comparision

Model 1 appears to have the lowest expected log pointwise predictive density (elpd) compared to Model 2 and Model 3, indicating potentially inferior overall predictive performance. According to the leave-one-out cross-validation (LOO) analysis, it exhibits the highest specificity and sensitivity rates. This discrepancy suggests a trade-off between predictive accuracy, as measured by elpd, and the ability of the model to accurately classify instances into true positive and true negative categories, as captured by specificity and sensitivity. While Model 1 may excel in correctly identifying true positive and true negative instances, it may sacrifice some predictive accuracy compared to Models 2 and 3. For our analysis of the impact of different factors on the mortality due to liver Cirrhosis. Model 2 is preferred because it's generalize our data more extensively.

7 Conclusion

In conclusion, our project "Liver Cirrhosis Bayesian Data Analysis" has provided valuable insights into the factors influencing the progression of liver cirrhosis using advanced Bayesian statistical methods. Through rigorous analysis of comprehensive datasets, we have explored the intricate interplay between various clinical, demographic, and laboratory parameters and the advancement of liver cirrhosis. Our hypothesis regarding the significant effects of key factors on liver cirrhosis status has been supported by our findings. Specifically, factors such as drug usage, age, sex, bilirubin levels, triglycerides levels, platelet count, and alkaline phosphatase levels have been identified as influential determinants of liver cirrhosis progression. Leveraging Bayesian data analysis methodologies has enabled us to quantify individual effects and elucidate potential interactions among these factors. Our interdisciplinary approach underscores the importance of collaboration between clinicians, statisticians, and data scientists in tackling complex medical challenges. Furthermore, our analysis of all model performances indicates that while Model 1 may excel in correctly identifying true positive and true negative instances, it may sacrifice some predictive accuracy compared to Models 2 and 3. Moving forward, our project aims to contribute to personalized intervention and management

strategies for liver cirrhosis by providing insights into predictive capabilities and model specifications.

Bibliography

Baptiste Auguie. gridextra: Miscellaneous functions for "grid" graphics. https://CRAN.R-project.org/package=gridExtra, 2017. R package version 2.3.

Paul-Christian Bürkner. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01.

John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, 3rd edition, 2019. URL https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Jonah Gabry and Paul-Christian Bürkner. bayesplot: Plotting for bayesian models. https://mc-stan.org/bayesplot/, 2024. R package version 1.11.0.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition, 2013.

Christian Heumann, Michael Schomaker, and Shalabh Shalabh. *Introduction to Statistics and Data Analysis*. 2016. ISBN 978-3-319-46160-1. doi: 10.1007/978-3-319-46162-5.

Jerry L. Hintze and Ray D. Nelson. Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998. URL http://www.jstor.org/stable/2685478.

JoeBeachCapital. Cirrhosis patient survival prediction. https://www.kaggle.com/datasets/joebeachcapital/cirrhosis-patient-survival-predicti Accessed: insert date.

Måns Magnusson, Michael Riis Andersen, Johan Jonasson, and Aki Vehtari. Leave-one-out cross-validation for bayesian model comparison in large data. *Journal Name*, Volume(Issue):Page numbers, 2020.

Bernhard Nowok, Gillian M. Raab, and Chris Dibben. synthpop: Bespoke creation of synthetic data in r. *Journal of Statistical Software*, 74(11):1–26, 2016. doi: 10.18637/jss.v074.i11.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

Aki Vehtari, Jonah Gabry, Måns Magnusson, Yu Yao, Paul-Christian Bürkner, Toni Paananen, and Andrew Gelman. loo: Efficient leave-one-out cross-validation and waic for bayesian models. https://mc-stan.org/loo/, 2024. R package version 2.7.0.

Brani Vidakovic. *Statistics for Bioengineering Sciences*. Springer Texts in Statistics. Springer New York, NY, 1 edition, 2011. doi: 10.1007/978-1-4614-0394-4. Hardcover ISBN: 978-1-4614-0393-7, Softcover ISBN: 978-1-4939-5144-4, eBook ISBN: 978-1-4614-0394-4.

Taiyun Wei and Viliam Simko. An introduction to corrplot package. 11 2021.

Hadley Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 2011. URL https://www.jstatsoft.org/v40/i01/.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.